STATISTICAL ANALYSIS OF GEOGRAPHICAL DATA

AN INTRODUCTION

SIMON J. DADSON

WILEY

Statistical Analysis of Geographical Data

Statistical Analysis of Geographical Data

An Introduction

Simon J. Dadson

School of Geography and the Environment, University of Oxford, UK

WILEY

This edition first published 2017

Registered Offices
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Name: Dadson, Simon J.
Title: Statistical analysis of geographical data : an introduction / Simon J. Dadson.
Description: 1 edition. | Chichester, West Sussex : John Wiley & Sons, Inc., 2017. | Includes bibliographical references and index.
Identifiers: LCCN 2016043619 (print) | LCCN 2017004526 (ebook) | ISBN 9780470977033 (hardback) | ISBN 9780470977040 (paper) | ISBN 9781118525111 (pdf) | ISBN 9781118525142 (epub)
Subjects: LCSH: Geography–Statistical methods. | BISAC: SCIENCE / Earth Sciences / General.
Classification: LCC G70.3 .D35 2017 (print) | LCC G70.3 (ebook) | DDC 910.72/7–dc23
LC record available at https://lccn.loc.gov/2016043619

Cover image: © Joshi Daniel / EyeEm/GettyImages
Cover design: Wiley

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Quantitative reasoning is an essential part of the natural and social sciences and it is therefore vital that any aspiring geographer be equipped to perform quantitative analysis using statistics, either in their own work or to understand and critique that of others. This book is aimed specifically at first year undergraduates who need to develop a basic grounding in the quantitative techniques that will provide the foundation for their future geographical research. The reader is assumed to have nothing more than rusty GCSE mathematics. The clear practical importance of quantitative methods is emphasized through relevant geographical examples. As such, the book progresses through the basics of statistical analysis using clear and logical descriptions with ample use of intuitive diagrams and examples. Only when the student is fully comfortable with the basic concepts are more advanced techniques covered. In each section, the following format is employed: (i) an introductory presentation of the topic; (ii) a worked example; and (iii) a set of topical, geographically relevant exercises that the student may follow to probe their understanding and to help build confidence that they can tackle a wide range of problems. Use of the popular R statistical software is integrated within the text so that the reader can follow the calculations by hand whilst also learning how to perform them using industry-standard open source software. Files containing the data required to solve the worked examples are available at https://simondadson.org/statistical-analysis-of-geographical-data.

I am grateful for the guidance and wisdom of my own academic advisers: Barbara Kennedy, who sadly died before the

book was completed, Mike Church, and Niels Hovius. I am grateful to seven anonymous readers of the book's outline for their positive support for the idea. To that I must add further thanks owed to colleagues at Oxford University, the Centre for Ecology and Hydrology, and elsewhere for their help and encouragement during the writing of this book. Particular thanks go to Richard Field, Richard Bailey, Toby Marthews and Andrew Dansie who read earlier drafts of the manuscript and made many useful suggestions that have undoubtedly improved the style of the book. My special thanks go to the large number of undergraduate and graduate students who have read the chapters and worked through the exercises in this book. Of course, any remaining errors or ambiguities are my own and I would be most grateful to have them brought to my attention.

At Wiley, I owe a considerable debt of gratitude to Fiona Murphy for her encouragement to undertake this project, Rachael Ballard for commissioning the work, and to Lucy Sayer, Fiona Seymour, Audrie Tan, Ashmita Thomas Rajaprathapan, Wendy Harvey and Gunal Lakshmipathy for their diligence in helping to see the work through to completion.

Finally, I would like to thank my wife, Emma, and our two children, Sophie and Thomas, for their support throughout the process of writing this book and for their tolerance of the time it has taken. To them this book is dedicated.

Oxford, 2016 *Simon J. Dadson*

1

Dealing with data

STUDY OBJECTIVES

- Understand the nature and purpose of statistical analysis in geography.
- View statistical analysis as a means of thinking critically with quantitative information.
- Distinguish between the different types of geographical data and their uses and limitations.
- Understand the nature of measurement error and the need to account for error when making quantitative statements.
- Distinguish between accuracy and precision and to understand how to report the precision of geographical measurements.
- Appreciate the methodological limitations of statistical data analysis.

1.1 The role of statistics in geography

1.1.1 Why do geographers need to use statistics?

Statistical analysis involves the collection, analysis and presentation of numerical information. It involves establishing the degree to which numerical summaries about observations can be justified, and provides the basis for forming judgements from empirical data.

Statistical Analysis of Geographical Data: An Introduction, First Edition.
Simon J. Dadson.

Take the following media headlines, for example:

We know in the next 20 years the world population will increase to something like 8.3 billion people.
Sir John Beddington, UK Government Chief Scientist[1]
2010 hits global temperature high.
BBC News, 20th January 2011[2]

Each of these statements invites critical scrutiny. The reliability of their sources encourages us to take them seriously, but how do we know that they are correct? It is hard enough to try to predict what one human being will do in any particular year, let alone what several billion are going to do in the next 20 years. How were these predictions made? How was the rate of change of world population calculated? What were the assumptions? What does the author mean by 'something like'? The number 8.3 billion is quite a precise number: why didn't the author just say 8 billion or almost 10 billion?

Similarly, how do we know that 2010 is the *global* temperature high, when temperature is only measured at a small number of measuring stations? How would we go on to investigate whether anthropogenic warming caused the record-breaking temperature in 2010 or whether it was just a fluke?

Statistical analysis provides some of the tools that can answer some of these questions. This book introduces a set of techniques that allow you to make sure that the statistical statements that you make in your own work are based on a sound interpretation of the data that you collect.

There are four main reasons to use statistical techniques:

- to describe and measure the things that you observe;
- to characterize measurement error in your observations;
- to test hypotheses and theories;
- to predict and explain the relationships between variables.

[1] http://www.bbc.co.uk/news/science-environment-12249909.
[2] http://www.bbc.co.uk/news/science-environment-12241692.

1.2 About this book

One of the best ways to learn any mathematical skill is through repeated practice, so the approach taken in this book uses many examples. The presentation of each topic begins with an introduction to the theoretical principles: this is then followed by a worked example. Additional exercises are given to allow the reader to develop their understanding of the topics involved.

The use of computer packages is now common in statistical analysis in geography: it removes many of the tedious aspects of statistical calculation leaving the analyst to focus on experimental design, data collection, and interpretation. Nevertheless, it is essential to understand how the properties of the underlying data affect the value of the resulting statistics or the outcome of the test under evaluation.

Two kinds of computer software are referred to in this book. The more basic calculations can be performed using a spreadsheet such as Microsoft Excel. The advantages of Excel are that its user interface is well-known and it is almost universally available in university departments and on student computers. For more advanced analysis, and in situations where the user wishes to process large quantities of data automatically, more specialized statistical software is better. This book also refers to the open-source statistical package called 'R' which is freely available from http://www.r-project.org/. In addition to offering a comprehensive collection of well-documented statistical routines, the R software provides a scripting facility for automation of complex data analysis tasks and can produce publication-quality graphics.

1.3 Data and measurement error

1.3.1 Types of geographical data: nominal, ordinal, interval, and ratio

Four main types of data are of interest to geographers: nominal, ordinal, interval, and ratio. *Nominal data* are recorded using categories. For example, if you were to interview a group of people and record their gender, the resulting data would be on a nominal,

or categorical, scale. Similarly, if an ecologist were to categorize the plant species found in an area by counting the number of individual plants observed in different categories, the resulting dataset would be categorical, or nominal. The distinguishing property of nominal data is that the categories are simply names – they cannot be ranked relative to each other.

Observations recorded on an *ordinal scale* can be put into an order relative to one another. For example, a study in which countries are ranked by their popularity as tourist destinations would result in an ordinal dataset. A requirement here is that it is possible to identify whether one observation is larger or smaller than another, based on some measure defined by the analyst.

In contrast with nominal and ordinal scale data, *interval scale* data are measured on a continuous scale where the differences between different measurements are meaningful. A good example is air temperature, which can be measured to a degree of precision dictated by the quality of the thermometer being used, among other factors. Whilst it is possible to add and subtract interval scale data, they cannot be multiplied or divided. For example, it is correct to say that 30 degrees is 10 degrees hotter than 20 degrees, but it is not correct to say that 200 degrees is twice as hot as 100 degrees. This is because the Celsius temperature scale, like the Fahrenheit scale, has an arbitrarily defined origin.

Ratio scale data are similar to interval scale data but a true zero point is required, and multiplication and division are valid operations when dealing with ratio scale data. Mass is a good example: an adult with a mass of 70 kg is twice as heavy as a child with a mass of 35 kg. Temperature measured on the Kelvin scale, which has an absolute zero point, is also defined as a ratio scale measurement.

It is important from the outset of any investigation to be aware of the different types of geographical data that can be recorded, because some statistical techniques can only be applied to certain types of data. Whilst it is usually possible to convert interval data into ordinal or nominal data (e.g. rainfall values can be ranked or put into categories), it is not possible to make the conversion the other way around.

1.3.2 Spatial data types

Geographers collect data about many different subjects. Some geographical datasets have distinctly *spatial* components to them. In other words, they contain information about the location of a particular entity, or information about how a particular quantity varies across a region of interest. In many contexts, it is advantageous to collect information on the locations of objects in space, or to record details of the spatial relationships between entities. The two main types of spatial data that can be used are vector data and raster (or gridded) data. *Vector data* consist of information that is stored as a set of points that are connected to known locations in space (e.g. to represent towns, sampling locations, or places of interest). The points may be connected to form lines (e.g. to represent linear features such as roads, rivers and railways), and the lines may be connected to form polygons (e.g. to represent areas of different land cover, geological units, or administrative units).

The locations of points must be given with reference to a coordinate system which may be rectangular (i.e. given using eastings and northings in linear units such as metres), or spherical (i.e. given using latitudes and longitudes in angular units such as degrees), but which always requires the definition of unit vectors and a fixed point of origin. The most common spherical coordinate system is that of latitude and longitude, which measures points by their angular distance from an origin which is located at the equator (zero latitude) and the Greenwich meridian (zero longitude). Thus the latitude of Buckingham Palace in London, UK, is 0.14°W, 51.50°N indicating that it is 0.14 degrees west of Greenwich and 51.5 degrees north of the equator.

Whilst spherical coordinate systems are commonly used in aviation and marine navigation, and with the arrival of GPS, terrestrial navigation usually uses rectangular coordinate systems. In order to use rectangular coordinates, the spherical form of the Earth must be represented on a flat surface. This is achieved using a map projection. An example of a map projection that is used to obtain a rectangular coordinate system is the Great Britain National Grid, in which locations are defined in metres east and north of a fixed origin that is located to the south west of the Scilly Isles. Thus to give a grid reference for Buckingham

Palace as (529125, 179725) is to say that it lies at a point which is 529.150 km east of the origin and 179.750 km north of the origin.

To reduce the amount of information that must be transmitted in practical situations, grid references are typically given relative to a set of predefined 100 km squares. In situations where quoting distances to the nearest metre is not justified they are usually rounded to a more suitable level of precision. The grid reference above, for Buckingham Palace, might be rounded to the nearest 100 m and associated with the box TQ [which has its origin at (500000, 100000)] to give TQ 291 797, where two letters indicate the grid square, the first three digits indicate the easting and the last three digits indicate the northing.

Raster data are provided on a grid, where each grid square contains a number that represents the value of the data within that grid square. Almost any kind of data can be represented using a raster. Examples of data that are collected in raster format include many types of satellite image, and other datasets that are sampled at regular intervals (see Section 2.1.3). The technical process of specifying the location of the raster in space is identical to the process used to locate a point, described above. It is also necessary to specify the resolution of the raster (i.e. the spacing between grid points and the extent or size of the domain).

1.3.3 Measurement error, accuracy and precision

All measurements are subject to uncertainties. As an example, consider a geographer wishing to measure the velocity of a river. One way to do this is to use a stopwatch to measure the time it takes a float to travel a known distance. What are the uncertainties involved in this procedure? One source of error is the reaction time of the person using the stopwatch: they might be slow starting the watch, or fast stopping the watch, or vice versa. Since each possibility is equally likely, this kind of error is termed *random error*. One way to measure the amount of random error in a measurement is to repeat the procedure many times: sometimes the time will be underestimated, other times we will overestimate the time. By analysing the variability or spread in our results, we can get a good estimate of the amount of random

error in our observation. If the spread is small, we say that our measurement is precise; if the spread is large, our measurement is less precise. The term *precision* is used to describe the degree to which repeated observations of the same quantity are in agreement with each other.

What if the stopwatch was consistently slow? In this case, all of the times measured would be shorter than they ought to be and no amount of repetition would be able to detect this source of error. Such errors are referred to as *systematic* errors, because we consistently underestimate the time taken if the stopwatch is slow, and consistently overestimate the time taken if the stopwatch is fast. If the amount of systematic error is low, we refer to our measurements as accurate; if the amount of systematic error is high, our measurement is less accurate. The term *accuracy* is used to describe the degree to which a measured value of a quantity matches its true value. Statistical analysis offers few opportunities to detect systematic errors, because we do not usually know the true value of the measurement that is being made: it is up to the person measuring the data to reduce the amount of systematic error through careful design of field, lab, or survey procedures.

A typical graphical analogy used to illustrate the difference between accuracy and precision involves a set of archery targets (Figure 1.1). Here, the archer is subject to random errors due to the wind or the steadiness of their hand; and potential systematic errors due to the design of the bow and arrow and its sight. Note the important point that it is impossible to assess the precision of a single measurement using statistical techniques.

1.3.4 Reporting data and uncertainties

The most straightforward way to communicate error is to give the best estimate of the final answer and the range within which you are confident that the measurement falls. Taking the earlier example of measuring the velocity of a river, suppose that we measure the velocity several times, giving the following estimates (in metres per second, or m/s):

$$0.5, 0.4, 0.5, 0.6$$

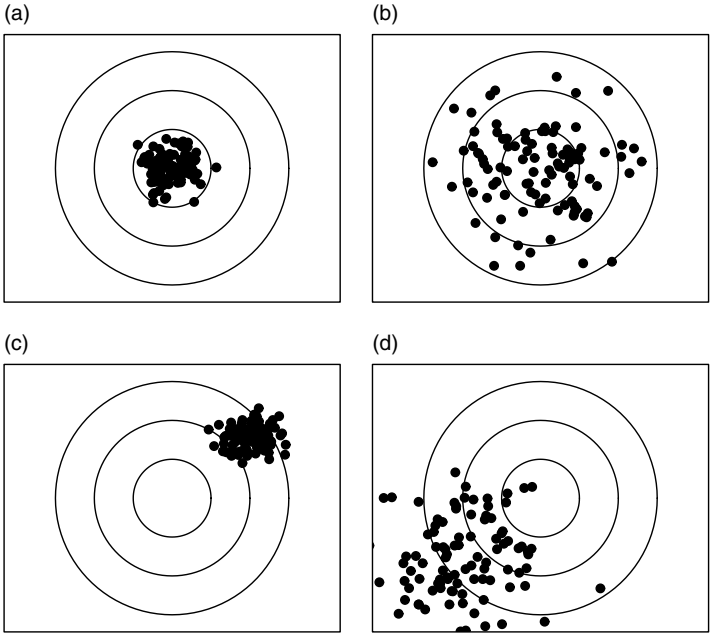

Figure 1.1 Accuracy and precision in archery. (a) High accuracy with high precision; (b) high accuracy with low precision; (c) low accuracy but high precision; (d) low accuracy and low precision. Note that without knowing the location of the archery target (i.e. the true value of the measured quantity), cases (c) and (d) are indistinguishable from (a) and (b), respectively.

The best estimate of the river's velocity is the average of these measurements, which is 0.5 m/s (calculated by adding up all of the values and dividing by the number of values). What about the range? At the most basic level, it makes sense to assume that the correct answer lies between the lowest (0.4 m/s) and the highest (0.6 m/s) values. We will want to refine this approach later on but for now we can say that the best estimate is $0.5\,\text{m/s} \pm 0.1\,\text{m/s}$, or to put it more generally:

$$\text{Measured value} = \text{best estimate} \pm \text{error} \tag{1.1}$$

Note that the '±' symbol is pronounced 'plus-or-minus'. In many situations, for data measured on a ratio scale, it is useful to

express the error as a percentage of the measured value. So 0.5 ± 0.1 m/s would become $0.5 \pm 20\%$ m/s.

It is worth noting that the example above, in which we used the range of observed values to estimate the error, is the simplest approach available. One of the aims of the statistical analyses described in this book is to provide more advanced ways to quantify this error, including the use of confidence intervals based on probabilities.

1.3.5 Significant figures

It is unwise to claim that your measurements are more precise than they really are. For example, it would be misleading to state that the average age of a group of interviewees was 23.357 if each person's age were known only to the nearest whole number. Rounding to an appropriate precision can be achieved by looking at the number of significant figures in the data.

The first significant figure (sig. fig. or s.f.) is the first figure which is not zero (reading from the left). For example, the first significant figure in the following numbers is indicated with a box:

1̲25, 0.01̲25, 0.00001̲25

The number of significant figures can then be counted from left to right, ignoring embedded zeros. The following examples all have four significant figures (boxed):

1̲2̲0̲5̲, 1̲2̲.̲0̲5̲, 0.1̲2̲0̲5̲, 0.001̲2̲0̲5̲

To round a number to:

1 s.f. look at the 2nd s.f.
2 s.f. look at the 3rd s.f.
3 s.f. look at the 4th s.f.

If the digit that you look at is less than 5, ignore it and round down. If it is 5 or more, round up by adding one to the digit in front.

You should round the final answer to reflect the precision of the original data. In general, the last significant figure quoted in your final result should be of the same order of magnitude as the error. It is important to remember that you should only round

the results of calculations as the last step in the chain of calculations. This is essential because otherwise you may incur round-off errors in the intermediate steps of the calculation.

1.3.6 Scientific notation (standard form)

Sometimes, it is necessary to use numbers that are very large or very small. For example, the amount of carbon stored in the terrestrial biosphere is enormous (approximately 600 000 000 000 000 000 g), whilst the size of a typical cell is tiny (about 0.00001 m). It is inconvenient to have to write down all these zeros. Scientific notation (or standard form) provides a concise way to report such numbers. The general definition of a number in scientific notation is: $a \times 10^n$, where a is a real number between 1 and 10, and n is an integer. So the numbers above convert to:

$$600\,000\,000\,000\,000\,000\,\text{grams} = 6.0 \times 10^{17}\,\text{g}\left(= 600\,\text{Pg}\right)$$

$$0.000\,01\,\text{m} = 1.0 \times 10^{-5}\,\text{m}\left(= 10\,\mu\text{m}\right)$$

The shorthand prefixes for various powers of 10 are given in Table 1.1.

Remember that the negative exponent refers to the reciprocal of the quantity concerned: $10^{-3} = 1/10^3$. This fact is sometimes used when writing down common units: so metres per second, or m/s, becomes m s^{-1}.

To convert a number into standard form, it is necessary first to write down the part of the number between 1 and 10 and then work out the power. For example, the number of people living in the UK is approximately 62 million people. To convert this to scientific notation, we have:

$$\begin{aligned} 62\,\text{million} &= 62\,000\,000 \\ &= 6.2 \times 10\,000\,000 \\ &= 6.2 \times 10^7 \end{aligned}$$

So, 62 million is 6.2×10^7. Note that computer programs and calculators often use a different way to display numbers in scientific notation, so in Microsoft Excel, for example, you might see 6.2×10^7 written as 6.2E + 07. Here the 'E' stands for '…times ten to the power of…'.

Table 1.1 Prefixes used to indicate powers of 10.

Number	Scientific notation	Prefix	Symbol
1 000 000 000 000 000 000	10^{18}	Exa-	E
1 000 000 000 000 000	10^{15}	Peta-	P
1 000 000 000 000	10^{12}	Tera-	T
1 000 000 000	10^{9}	Giga-	G
1 000 000	10^{6}	Mega-	M
1 000	10^{3}	Kilo-	k
1	1	—	—
$\frac{1}{1000}$	10^{-3}	Milli-	m
$\frac{1}{1000000}$	10^{-6}	Micro-	μ
$\frac{1}{1000000000}$	10^{-9}	Nano-	n
$\frac{1}{1000000000000}$	10^{-12}	Pico-	p
$\frac{1}{1000000000000000}$	10^{-15}	Femto-	f
$\frac{1}{1000000000000000000}$	10^{-18}	Atto-	a

1.3.7 Calculations in scientific notation

Calculations using scientific notation are sometimes easier than with ordinary numbers, especially if the quantities are particularly large or small. The basic procedure is to group the numbers and the powers of 10, then work out the multiplications or divisions using the laws of indices (i.e. multiplication requires the addition of indices; division requires their subtraction).

Worked Example 1.1 The population of England and Wales in the 2011 census was estimated to be 56 075 900 people. The total land area of England and Wales is 245 000 km^2. Express each of these figures in standard form retaining all their precision and then find the population density in England and Wales to 3 significant figures.

Solution First, to convert the figures into standard form (scientific notation) and round them to 3 s.f., the population becomes 56.0759×10^6 (or 56.0759 million) and the total land area is 2.45×10^5 km^2. To divide the population by the land area we write that:

$$\begin{aligned}\text{Population density} &= \left(56.0759 \times 10^6\right) \div \left(2.45 \times 10^5\right) \\ &= \left(56.0759 \div 2.45\right) \times \left(10^6 \div 10^5\right) \\ &= 22.8881 \times 10^1 \\ &= 229 \text{ people per square kilometre, to 3 s.f.}\end{aligned}$$

Exercises

1 The population of Northern Ireland in the 2011 census was 1 810 900 and the land area of Northern Ireland is 13 840 km^2. Express these numbers in scientific notation to 3 significant figures and calculate the population density in Northern Ireland.

2 Define the terms accuracy and precision and explain how you would quantify them in a geographical study of your choice.

3 The average annual rainfall in the catchment of the Thames to Kingston is 720 mm per year. The area of the Thames catchment draining to this point is 9948 km^2. The average discharge of the Thames at Kingston is 78 m^3/s. What fraction of the rain falling in the Thames catchment travels to the gauging station at Kingston?

2

Collecting and summarizing data

STUDY OBJECTIVES

- Appreciate the range of possible sampling procedures and the importance of randomization and replication in research design.
- Recognize a range of different graphical methods for presenting data (e.g. histograms, time series, scatter plots) and understand the circumstances in which each can be used.
- Understand the range of measures of central tendency: mean, median and mode.
- Appreciate some measures of dispersion: variance, standard deviation and inter-quartile range.

2.1 Sampling methods

2.1.1 Research design

In statistics the entire group of entities that is of interest is called the population, and it is desirable to be able to make statements about the population from a smaller fraction of the population, which is called a sample. Examples of geographical research in which sampling techniques are typically used include population census surveys, assessments of biodiversity from field samples, monitoring of atmospheric and oceanic processes using sparsely deployed instruments, and surveys and questionnaires designed to support interviews.

It is clear that in any study the results will be applicable only to the measurements made in that study, although it might be

Statistical Analysis of Geographical Data: An Introduction, First Edition.
Simon J. Dadson.